Технология стандартизации дидактических тестов

Михайлычев Евгений Аркадьевич — зав. кафедрой общей педагогики Таганрогского государственного педагогического института, научный руководитель Центра педагогических измерений Ростовского областного института повышения квалификации и переподготовки работников образования, профессор, доктор педагогических наук.

Стандартизация дидактических тестов (как и других методик) представляет собой длительный процесс экспериментального выявления диагностических возможностей теста как инструмента познания. Мы рассмотрим его логическую последовательность, хотя в некоторых случаях отдельные процедуры стандартизации могут и синхронизироваться (особенно при наличии группы высококвалифицированных разработчиков).

Обеспечение содержательной валидности тестов

Необходимость содержательной валидности признаётся фактически всеми, кто работает над проблемами стандартизации методик и обеспечением достоверности экспериментальных данных. Под **содержательной валидизацией** дидактического теста понимается совокупность последовательных процедур теоретического анализа и опытно-экспериментальных работ. В процессе их проведения предстоит с помощью различных методов определить, в какой степени создаваемый или проходящий экспертизу тест в целом и составляющие его тестовые задания соответствуют целям обучения и содержанию учебного предмета, успешность усвоения которых диагностируется.

Предлагаемая нами методика содержательной валидизации разработана опытно-экспериментальным путём. Она последовательно описывает процедуры содержательной валидизации создаваемого дидактического теста.

При оценивании **содержательной валидности эксперты** устанавливают, насколько **задания** теста соответствуют **содержанию предмета**.

Содержательная валидность теста имеет, в нашем представлении, четыре взаимосвязанных аспекта:

• **куррикулярную**, или **программную** валидность (адекватность теста учебному плану, программе и т.д.);

• **конструктную** (его адекватность структуре, модели-конструкту диагностируемой области знания или умениям);

• **концептуальную** (отражение в тесте доминирующих концепций, трактующих так или иначе диагностируемую систему знаний и умений);

• **психологическую** (адек-ватность отражения в тесте развивающих целей и задач, лежащих в основе изучения учащимся диагностируемого материала).

Куррикулярная валидность предметного теста

Термин *куррикулярная валидность* применяется для характеристики соответствий теста по предмету, учебному плану и программе, для измерения уровня усвоения которых разработан тест. Возможно исследование куррикулярной валидности созданного или адаптируемого (проходящего экспертизу) предметного дидактического теста по пяти позициям:

• на соответствие теста содержанию типовой программе курса (если таковая имеется);

• на соответствие рабочей программе учебного курса, предмета (если она общая для учащихся различных специальностей или профилей подготовки, изучающих курс);

• на соответствие теста рабочим программам по отдельным специальностям или типам образовательных учреждений, если при изучении курса в этих программах имеются существенные различия в их структуре, расположении учебного материала в хронологическом порядке, в выделенных для изучения разделов часах, в номенклатуре и содержании лабораторно-практических работ, семинарских занятий. Этот вариант больше подходит для школ с

разными уклонами, колледжей, вузов — особенно многопрофильных;

• на соответствие теста базовому учебнику по курсу;

• на соответствие теста учебным пособиям по курсу в случае, когда признанного базового учебника нет или обучение ведётся по экспериментальной программе, по авторскому учебному пособию, проходящему апробацию на ограниченном контингенте учащихся.

• на соответствие теста собственно учебному и стандарту образования, в том виде, как он разрабатывается в России (проводится только в тех случаях, когда разработан или адаптируется межпредметный (междисциплинарный) дидактический тест. Например, для поступления в вуз по дисциплинам естественнонаучного цикла, для определения естественнонаучной или гуманитарной подготовки после окончания первых двух курсов вуза как первой ступени подготовки специалиста и т.д.).

При сопоставлении результатов образовательной деятельности государственных и альтернативных образовательных учреждений ориентация теста на базовый учебный план и стандарт образования позволит при рубежном и итоговом контроле определять «кто есть кто», реализованы или нет требования Закона «Об образовании» и в какой степени.

В зависимости от конкретной спецификации и имеющихся резервов времени разработчика (либо эксперта), проверяющего куррикулярную валидизацию, из приведённой выше программы максимум выбираются оптимальные варианты (например, 2 и 4), без которых трудно обойтись. Остальные же варианты валидизации проводятся в более позднее время для дополнения представлений о тесте (если разработчик или пользователь желают добросовестно сделать своё дело).

Возможны также следующие **ситуации куррикулярной валидизации дидактического предметного теста:**

• А. Тест находится в стадии замысла, как система ещё не создан (имеются цели и, возможно, подобраны определённые задания).

• Б. Тест разработан, несколько раз проводился, имеются какие-то обработанные результаты, но проверки надёжности и валидизации не проходил.

• С. Имеется тест, прошедший стандартизацию (но на другом контингенте, возможно, по другим рабочим программам), который требует дополнительной проверки. Например, турецкие или американские тесты из-за различия учебных планов, программ и учебников все должны пройти повторную валидизацию.

Поскольку основные процедуры содержательной валидизации имеют много общего, а главная проблема многоэтапного контроля в вузах состоит в создании своих республиканских стандартизированных тестов, особое внимание уделяется разработке нового предметного теста.

Процедуры куррикулярной валидизации

Во-первых, необходимо сделать иерархическую нумерацию информации.

В типовой или рабочей программе, в учебном пособии последовательно проставляется иерархическая нумерация:

а) разделов;

б) тем;

в) подтем (глав);

г) рассматриваемых вопросов, имеющих относительную самостоятельность (пункты программы, смысловые группы учебных материалов, обычно параграфов — в учебниках и пособиях);

д) учебных элементов.

Так как в типовых и рабочих программах до уровня учебных элементов детализация обычно не доходит, это делается на основе учебника, учебного пособия, а в программах — на основе планов и методразработок лабораторно — практических, семинарских занятий (см. форма 1).

Во-вторых, надо подготовить бланк экспертизы. Для экспертного анализа — основного

звена куррикулярной валидизации — составляется и распечатывается (в количестве, соответствующем числу экспертов плюс 2 — 3 экземпляра) «Бланк экспертной оценки значимости (процентного веса) учебного материала» (форма1). В левой колонке приводится сквозная нумерация оцениваемых позиций учебного материала (в соответствии с типовой или рабочей программой либо учебником, учебным пособием). В прочих колонках указываются параметры, по которым проводится экспертное оценивание. Рекомендуется выделять следующие параметры, соответствующие различным иерархическим уровням экспертного оценивания:

1-й уровень — отношение раздела к теме;

2-й уровень — отношение главы к разделу;

3-й уровень — отношение темы к главе;

4-й уровень — отношнение вопроса к теме;

5-й уровень — отношение учебного элемента к вопросу.

Это не все возможные варианты экспертного оценивания, но практика показывает, что больше загружать экспертов не стоит, так как на основе этих данных можно провести все необходимые перерасчёты, используя математические познания на уровне школьника 6-го класса. В вузе при выполнении этой работы можно подключить группы экспертов — ведущих преподавателей кафедр, которые продолжают специальную подготовку студентов, основанную на знаниях из той предметной области, для которой создаётся тест. Учёт их позиции как пользователей «продукта» ранее работавших преподавателей других дисциплин этого же предметного цикла позволит внести коррективы и в рабочие программы, и в структуру теста. Это необходимо, если с его помощью предлагается решать задачи селекции по специальностям, уровням и профилям подготовки, а в вузе или колледже — отбора на стажировку или доучивание за границей и т. п.

В-третьих, необходимо дать инструкцию экспертам. Эксперты обеспечиваются инструкциями примерно такого содержания (см. форма 1):

Форма 1. (колонки таблицы — *прим. сост. эл. версии*)

№

Содержание материала

1-й уровень — отношение раздела к курсу (курс 100%)

2-й уровень — отношение главы к разделу (раздел 100%)

3-й уровень — отношение темы к главе (глава 100%)

4-й уровень — отношение вопроса к теме (тема 100%)

5-й уровень — отношение учебного элемента к вопросу (100%)

«Уважаемый коллега! Убедительно просим Вас принять участие в качестве эксперта в стандартизации предметного (многоэтапного) дидактического теста по курсу ....., создаваемого в ..... для ..... Если Вы согласны, Вам предлагается следующий алгоритм выполнения работы:

1. В 3-й колонке слева (отношение раздела к курсу) оцените в процентах (без десятых), как Вы лично представляете процентную значимость каждого раздела в общей структуре курса (вариант для модификации теста: «по специальностям …….»), например, 1-й раздел — 20%, 2-й — 10% и т. д.

2. В 4-й колонке слева, принимая за 100% объём информации в каждом разделе, проставьте таким же образом, как Вы оцениваете процентный вес (значимость) каждой главы.

3. В 5-й колонке та же самая процедура проводится относительно процентного веса каждой темы в структуре главы (глава 100%).

4. В 6-й колонке та же процедура, но за 100% принимается тема.

5. В 7-й колонке та же процедура, но за 100% принимается каждый относительно самостоятельный вопрос одной темы. Учебным элементом можно при этом считать каждую самостоятельную формулу, формулировку правила или закона, новые для учащихся термины, график или таблицу и т. п. Заранее благодарим Вас за помощь в нашем исследовании!»

В-четвертых, следует продумать проведение самой процедуры экспертизы, включая завершающий её подсчёт и суммирование по средним арифметическим оценок группы экспертов (или нескольких экспертных групп), а также последующие после работы экспертов

дополнительные операции разработчика с данными экспертизы.

В зависимости от того, для какого уровня многоэтапного контроля знаний создаётся данный тест, в большей или в меньшей степени используется объём полученной информации. Каковы задачи разработчика теста?

1. Просчитать средние оценки как по каждой группе экспертов по всем позициям, так и желательно по подгруппам (если число экспертов в каждой из них достигает хотя бы 8 — 10 человек): хорошо бы учесть средние квадратические отклонения и вариационный размах оценок.

2. В специальной таблице (форма 2) пересчитать на основе средних оценок экспертов их представления по позициям: A — отношение темы к разделу; B — отношение темы к курсу; C — отношение вопроса к теме, D — отношение вопроса к главе; E — отношение вопроса к разделу; F — отношение вопроса к курсу. Внести в таблицу C — отношение вопроса к теме (указанное экспертами при подсчёте в форме 1 в колонке 4).

Форма 2. Сравнительная значимость учебного материала по рабочей программе и по контрольным вопросам теста

По типовой программе, учебнику			**По контрольному вопросу теста**			
3 уровень	**A**	**B**	**C**	**D**	**E**	**F**
15	1.8	0.29	20.3	3.0	0.36	0.06

При построении заданий для тестов рубежного и итогового контроля во многих случаях такой уровень детализации окажется вполне достаточным для формулировки отдельных тестовых заданий. В тех же случаях, когда по каким-либо принципиальным соображениям, связанным с особой значимостью отдельных учебных элементов, возникнет необходимость выделить на этот элемент отдельные самостоятельные тестовые задания, целесообразно просчитать и его процентный вес (значимость) по всем уровням иерархии учебного материала.

В результате будут получены процентные веса для каждого планируемого задания теста. Если же имеется уже готовый тест, то после его спецификации можно будет на основе полученных данных легко определить, каков реальный процентный вес каждого тестового задания (и всей их совокупности) в контроле знаний по предмету именно в данном тесте (в сравнении с реальным процентным весом данного вопроса, найденным по полученной таблице экспертной оценки значимости учебного материала). И что этот тест на самом деле диагностирует, в какую сторону (фактология, закономерности, формулы и т.п.) в нём есть «перекосы», в чём его доминанта диагностики, сильные и слабые стороны.

Затем рассчитываются процентные веса структурных компонентов учебного материала.

Как «работает» форма 2, можно проследить на примере из первого «преподавательского» варианта теста по курсу общей физики Бухарского технологического института, который подвергался переработке. В варианте теста (субтеста) рубежного контроля был сформулирован вопрос-задание: «Что называется средней скоростью движения?» (задание 2-го уровня по типологии В.П. Беспалько) наряду с другими тремя вопросами-заданиями.

По экспертным оценкам его реальный процентный вес указан по отношению к разделу, знания которого диагностировались в субтесте 0,36%. Но весь субтест включал 25 вопросов. Значит, при условии, что все вопросы имеют одинаковую значимость (а на самом деле это не так), задание должно было охватывать 4% объёма материала (почти в 10 раз больше). Надо либо увеличить количество заданий такого уровня обобщённости в тесте, либо снять это задание как мало информативное и заменить его таким, которое (при сохранении в тесте 25 заданий) будет охватывать значительно больший объём учебного материала.

Совершенно ясно также, что в итоговый (курсовой) тест задание №1 в такой формулировке не может входить, ибо получило лишь 0,06% «значимости», а тест вряд ли будет включать более 1000 заданий. На каком же уровне оно может работать в такой формулировке, без изменений? Форма 2 точно подсказывает: при контроле знаний по теме и максимум — по главе. От задания не отказываются, оно просто «переводится» в тест более низкого ранга—

текущего контроля знаний.

3. Определение количества заданий в тесте (субтесте). В опыте тестологии нет строгих ограничений на количество заданий дидактических тестов: оно доходит в стандартизированных американских тестах до 200 и более, не ограничивается каким-то круглым, удобным для подсчёта числом (особенно с появлением и применением для тестирования компьютерной техники). Заданий нужно столько, сколько потребуется для того, чтобы по возможности полно отразить основное содержание диагностируемого объёма знаний. Если отдельные вопросы по данным экспертизы набирают очень небольшие «процентные веса значимости», то их можно использовать в блочных тестовых заданиях с более сложной композицией. Только «сырые баллы», получаемые за них, окажутся пропорционально меньше. Никто не обязывает разработчика включать примитивные задания в тест для того, чтобы каждое из них давало обследуемому по одному баллу. За счёт совершенствования техники композиции теста можно даже самые незначительные (по процентным весам) учебные элементы включать в структуру более сложных по форме заданий и учитывать их усвоение при контроле знаний.

Коэффициенты измерения куррикулярной валидности

Вопрос о единицах измерения куррикулярной валидности в психодиагностической литературе не рассматривался, поскольку ранее содержательная валидизация оценивалась как процедура качественного, теоретического анализа. Предложенная нами методика, применяемая к созданному и экспертируемому дидактическому тесту, позволяет в определённой степени использовать количественные меры. Валидность близкого к идеальному предметного дидактического теста будет тяготеть к единице (1,0), никогда её реально не достигая. Только теоретически возможно построить дидактический тест, полностью и адекватно проверяющий не только все вопросы курса, но и знание всех учебных элементов и даже более мелких дидактических единиц структуры элементов (например, всех символов, входящих в формулу, всех особенностей приёмов построения графиков, таблиц).

Поскольку в дидактическом тесте самостоятельным подструктурным компонентом является тестовое задание, то, учитывая данные, получаемые в форме 2, можно предложить следующие коэффициенты куррикулярной валидности:

а) коэффициент полноты охвата содержания программ тестовыми заданиями — по сумме процентных весов основных тестовых заданий:

$F_C = \Sigma(F_{C1}, F_{C2},... F_{Cn})$,

где F_C — тестовый коэффициент куррикулярной валидности, $F_{C1}...F_{Cn}$ — коэффициент 1-го и n-го заданий, n — число заданий, (c — curriculum);

б) коэффициент полноты охвата содержания разделов курса — вычисляется аналогичным образом по позиции E как E_C при экспертизе хода рубежного контроля. В тесте итогового контроля предстаёт в виде формулы:

$F_C = \Sigma(F_{C1}, F_{C2},... F_{Cm})$,

где m — количество разделов теста;

в) аналогичным образом могут быть подсчитаны коэффициенты полноты охвата главы (целесообразен при оценке куррикулярной валидности учебника, учебного пособия):

$D_C = \Sigma(D_{C1}, ... D_{Cn})$,

где n — количество глав в структуре книги, и

г) коэффициенты охвата темы:

$C_C = \Sigma(C_{C1}, ... C_{Cn}) / n$

где n — количество тем.

Для перестройки отдельных заданий целесообразен и коэффициент полноты отражения вопроса в j-м тестовом задании, но его можно вывести только в том случае, если экспертиза детализирована до уровня оценки процентного веса отдельных элементов к вопросу субтеста (5-й уровень), на что эксперты из-за большого объёма работы идут крайне неохотно.

Однако если в отдельных случаях удаётся получить такие экспертные оценки, то применительно к отдельному тестовому заданию его можно просчитать по формуле:

TQ1=E(*de*1+*de*2 + ... *den*),

где TQ1 — тестовый вопрос №1 (от test-question), *de*1, *den* — процентные веса отдельных учебных элементов в вопросе— от 1-го до *n*-го (от didactical element — англ.), а *n*— количество учебных элементов, отражённых в тестовом задании.

Наиболее точные подсчёты куррикулярной валидности будут проводиться с учётом TQ1 как исходного элемента последовательного «восхождения» от 5-го уровня экспертизы до 2-го или даже 1-го. Применение всей этой совокупности приведённых коэффициентов при экспертном оценивании по форме 1 (содержания программы учебника) и в форме 2 (реальный тест, вновь созданный или адаптируемый) разработчик или эксперт получит по тесту следующую информацию:

а) по куррикулярной адекватности каждого тестового задания; б) по адекватности отражения темы в заданиях теста; в) по адекватности отражения главы в заданиях теста; г) по адекватности отражения раздела в заданиях; д) по адекватности отражения курса в заданиях данного теста.

Наибольшая полнота информации достигается при экспертизе, доведённой до уровня учебных элементов, что позволяет при расчёте каждого более высокого иерархического уровня в системе «вопрос — тема-глава — раздел — курс» использовать как исходные более точные «процентные веса» значимости каждого задания. При подсчёте среднеквадратических отклонений и других принятых статистических мер возможно предусмотреть с достаточно высокой точностью вариацию разброса данных оценивания.

В любом случае, даже при более «грубом», но более вероятном на практике, оценивании учебной программы (учебника, пособия) и теста несколькими экспертами с простым арифметическим усреднением их оценок разработчик или тот, кто проводит адаптацию теста, будет иметь достаточно выраженную картину куррикулярной валидности теста. Это намного лучше, чем иметь только иллюзию того, что «тест хороший, потому что он нравится мне» (или прислан из авторитетного вуза и т.д.).

Методика конструктивной валидизации дидактического предметного теста

Конструктная валидность дидактического валидного теста — это степень отражения в нём теоретической модели курса — его структурных пропорций и основных компонентов на всех уровнях, вплоть до «низшего» уровня учебных элементов.

Куррикулярная валидность даёт нам лишь основу для определения конструктной валидности. Она показывает только одно: отражены ли в тесте учебные элементы (и какие именно), вопросы, темы, главы, разделы.

Но логическая структура учебного материала — это не просто совокупность элементов знания, но и их определённая последовательность, со сложной системой координационных и иерархических связей.

На базе куррикулярной валидности намного легче проводить конструктную валидизацию. Она содержит следующие процедуры при разработке нового дидактического теста:

1. В текстовой и графической форме представляется модель структуры знаний и умений в виде структурно-логической схемы с иерархическими и координационными связями.

2. Строится аналогичная и по возможности адекватная ей модель теста (желательно — как структурно-логическая схема). Если же проводится адаптация уже готового (например, переводного) теста, то на основе спецификации этого теста также строится структурно-логическая схема его конструкта. А затем она соотносится с заявленной разработчиками теста структурной схемой (конструктом) учебного материала.

Конструктная содержательная валидность описывается с точки зрения того, какие элементы модели в большей или меньшей степени отражены в тесте, а какие — не отражены, позволяет ли тест (через наличие и взаимосвязь заданий) диагностировать внутрипредметные связи.

В принципе возможно математическое описание конструктной валидности дидактического теста, но её технология не разработана.

Концептуальная валидизация дидактического теста

В основе каждого учебника, учебного пособия, любой типовой или рабочей программы всегда объективно заложена определённая концепция учебного предмета. Она отражает представления авторов о том, что в учебном материале является более или менее значимым, какие разделы, темы, главы, вопросы, учебные элементы заслуживают в изложении, объяснении большего внимания, а какие знания должны быть закреплены особо прочно, какие— быть «информационным фоном».

Далеко не всегда (к сожалению) эта концепция предмета чётко и последовательно, хотя бы кратко, описывается самими авторами, так как далеко не всегда она осознаётся самими авторами теста на методологическом уровне. Тем не менее данные куррикулярной и конструктной валидизации теста (как создаваемого, так и экспертируемого, например, переводного) могут дать необходимую информации о том, какой дидактической или предметно-методической концепции по тому или иному разделу (главе, теме) теста реально придерживается разработчик.

Не исключена ситуация, когда цели изучения учебного предмета, его рабочая программа и базовый учебник направлены на формирование умений решать теоретические и прикладные задачи, а тест проверяет только прочность запоминания правил, законов и формул. Задачи на их применение не ставятся, хотя все ключевые элементы конструкта и содержания отражены, но только на уровне воспроизводства имеющихся знаний.

Другой вариант — в учебник включено немало новейшего материала, характеризующего современные достижения науки, который не нашёл отражения в тестовых заданиях или представлен примитивными заданиями на распознавание.

Соотнесение целей обучения и реальных (а не провозглашаемых) целей тестирования позволяет выявлять такого рода противоречия. При разработке нового теста анализ его концептуальной валидности позволяет привести его в соответствие с концепцией предмета и учебника, а при экспертизе адаптируемого теста решить вопрос о целесообразности его использования.

Количественные методы при описании концептуальной валидности теста пока не применяются, она строится целиком и полностью на дидактическом и логико-психологическом анализе заданий, основанном на результате спецификации, куррикулярной и конструктной валидности. Это, однако, вовсе не означает, что математические подсчёты здесь нельзя применять — просто, как отмечает Анна Анастази, многие разработчики тестов идут самым простым путём*. Современный математический аппарат моделирования позволяет решать и более сложные задачи, как, например, описание экономических моделей. Необходимы специальные исследования в этом направлении, в процессе которых, как нам кажется, может быть создана методика математического описания концептуальной валидности (особенно при опоре на контент-анализ).

---

* Анастази А. Психологическое тестирование. М.: Педагогика, 1982.

С формальной стороны при анализе конструктной и концептуальной валидности теста можно составить перечень отклонений теста от типовой (учебной) программы, от конструкта учебника с указанием рекомендаций по совершенствованию, доводке текста.

Такой объём проделанной работы позволит, на наш взгляд, гарантировать разностороннюю **содержательную валидизацию любого дидактического теста.**

Спецификация теста

Спецификация является, по признанию ведущих текстологов (А. Анастази — США, Г. Витцлак — Германия, А.Г. Шмелев, В.С. Аванесов — Россия), обязательной операцией эмпирического анализа адаптируемого теста или условием создания нового содержательно валидного теста. Спецификация заключается в заполнении таблицы, в которой указываются: а) номера вопросов (тест-заданий); б) что каждое из них конкретно диагностирует (в операциональных понятиях).

Чем конкретнее сформулированы диагностические цели и задачи теста в целом, тем легче определить диагностические задачи по каждому тестовому заданию (а таких задач может быть несколько для одного задания). Спектр возможных диагностических задач отражён в номенклатуре видов тестов и тестовых заданий разного типа.

Спецификация позволяет видеть главную дидактико-диагностическую цель постановки каждого задания и необходима (хотя бы в упрощённой форме) для **начальной** стадии разработки дидактического теста или для его качественной экспертизы.

А. Анастази показывает другой распространённый в практике тестологов США подход к форме спецификации на примере использования при этой процедуре одновременно двух параметров: изучаемого содержания и учебных целей. В этом случае создавалась таблица, в которой в левой большой колонке перечислялись темы или содержательные категории, которые должны были быть охвачены тестом. В верхней части таблицы приводились четыре тестируемые цели обучения. На пересечении строк и столбцов отмечалось, сколько заданий должны отражать соотношение и взаимосвязь содержания и целей обучения*.

* Анастази А. Психологическое тестирование. С. 50–51.

Этот подход имеет то преимущество (по сравнению с представленным в форме 3), что позволяет соотносить дидактические и психологические цели и тем самым быть эмпирической основой и психологической, и конструктной валидизации теста. Такой приём целесообразен для итоговой спецификации теста, а также при анализе уже имеющегося дидактического теста для его корректировки и сопоставления с другим тестом, сравнения дублирующих форм и т.д.

Форма 3. Упрощённая форма спецификации дидактического теста

№	Что диагностирует
1	Знание формулы А
2	Прочность усвоения методики расчёта ... по формуле А
3	Понимание сферы применения формулы А в предметной области знания по ... (физики, химии и т.п.)
4	Понимание значения символов, включённых в формулу В и т. д.

Экспертное оценивание сложности проектируемых заданий

Поскольку, как известно, тестовые задания объективно обладают определённой степенью сложности, вызванной содержательным аспектом изучаемого материала, целесообразно ещё на стадии разработки теста попытаться её определить. Наиболее эффективным средством здесь является экспертное оценивание с привлечением тех же экспертов, что и при куррикулярной валидизации теста.

Полученные данные должны помочь в решении таких исследовательских задач:

1) перепроверить экспертные оценки процентного веса учебных элементов, вопросов, тем, глав, разделов на материале анализа заданий готового варианта теста до его проведения;

2) прогнозировать **трудность заданий** теста для всего контингента обследуемых учащихся. Сопоставление предварительных оценок экспертов прогнозируемой степени сложности заданий и результата тестирования по всей выборке по её подгруппам позволяет одновременно получить сведения о прогностических способностях экспертов и сравнить взаимосвязь сложности заданий (как относительно объективного фактора) с их трудностью для учащихся с различной успеваемостью, биографо-демографическими характеристиками (полученным методом анализа этих данных по личным делам учащихся);

3) обосновать и ввести в итоговую оценку шкалу перевода «сырых» тестовых баллов в оценочные, поскольку различная степень сложности заданий объективно требует столь же дифференцированного подхода к их оцениванию.

Форма проведения этой процедуры проста: экспертам предлагается вариант подготовленного теста и они около каждого задания должны по 10-балльной шкале оценить предлагаемую ими степень сложности для студентов. Затем данные суммируются, подсчитываются

среднеарифметические. При увеличении количества экспертов (до 20 и более) можно просчитывать среднеквадратические отклонения, моду и вариационный размах в оценках экспертов по каждому тестовому заданию.

Сопоставление данных экспертизы с результатами тестирования по подгруппам, а также по академическим группам, в которых проводили занятия сами эксперты (если в их роли выступали педагоги этого же образовательного учреждения), должны были дать информацию, значимую для корректировки теста и совершенствования экспертных методик.

Психологическая валидизация дидактического теста логически завершает описанные процедуры, хотя может проводиться и автономно. Суть её заключается в выявлении направленности теста (его субтестов и заданий) на выполнение обследуемыми определённых психических (а в тестах действия — и сенсомоторных) действий и операций.

Известно, что все стандарты образования и учебные программы (а при профподготовке— профессиограммы и психограммы) имеют определённые цели **развития личности** и формирования связанных с ними психических свойств (логического мышления, творческого воображения, механической и смысловой памяти, сенсомоторных характеристик). Насколько тест нацелен на проверку достижения учащимися указанных целей психического развития, может показать только психологический анализ. Такой анализ должен опираться на известные в психологии классификации психических качеств, конкретную концепцию структуры. Учитывая имеющиеся различия в трактовках, надо указывать, чью концепцию мы используем при психологическом анализе теста.

Независимо от устремлений разработчика при создании теста, тест всегда требует какой-то определённой реакции и психических действий испытуемого. Вполне знакома многим ситуация, когда тест, особенно **гомогенный** по форме предъявления знаний, полностью нацелен на элементарные психические операции и механическую память (в то время как целями диагностики авторы провозглашали выявление сообразительности, логичности мышления и т.д.). На основе анализа **психологической** валидности можно в каждом дидактическом тесте создать **дополнительные** шкалы, демонстрирующие, какие «укрупнённые» психологические блоки качеств тест дополнительно выявляет.

В межпредметных и «надпредметных» по своему характеру тестах обучаемости психологическая валидизация более тесно интегрируется с **конструктной** и **концептуальной**, почти вытесняя **куррикулярную**.

Психологическая валидность проявляется при выполнении операций **спецификации** теста в полном объёме: как рекомендует А. Анастази, когда в таблице будут соотноситься предметные области знаний (требования стандартов) с дидактико-диагностическими целями тестирования.

С содержательной валидностью связан ещё один вид валидности, который, по сути, валидностью не является.

Это — **очевидная валидность**. Она относится не к тому, что тест на самом деле измеряет, а к тому, что он при первом рассмотрении якобы измеряет (каким он воспринимается разработчиком, пользователем-диагностом, самим испытуемым, официальным лицом, принимающим решение о тестировании). Тем не менее нередко именно очевидная валидность является решающей в выборе методик малоквалифицированными диагностами.

Инструктивно-методическое обеспечение диагноста и испытуемого

Стандартизированные дидактические тесты обязательно имеют «Руководство к тесту», содержащим основные сведения о самом тесте и рекомендации по организации процедур тестирования, анализу и обработке данных: полное название теста, характеристика его диагностических целей, задач, сведения об авторах-разработчиках: кто, где работает, учёные степени и звания (иногда — чем известны в науке, методики, книги, должности в престижных профессиональных или национальных, международных организациях). Здесь обязательно даётся описание структуры теста (иногда и его **спецификация**), сведения, имеющиеся у из-

дателя теста, его модернизации (модификациях, версиях). В руководстве должны быть сведения о репрезентативности выборки апробации, её сильных и слабых сторонах (какие подгруппы проектируемого контингента обследуемых отражены хуже или лучше). Необходимы данные о надёжности и методах её проверки (желательно не только по всей выборке, но и по подгруппам обследованного контингента), а также сведения о валидности — содержательной (относительно каких учебных программ или каких укрупнённых дидактических целей, какие показатели соответствуют каким внешним критериям), локальной — для каких подгрупп обследованных, конкурентной — с какими другими тестами и методиками.

В руководстве к тесту должны быть **нормативы** (с описанием процедур их выработки или даже с приложением исходных таблиц, обязательно— со ссылками на литературу, в которой отражены методологические исследования по данному тесту).

Нормативы в хороших стандартизированных тестах даются не только по всему контингенту (обычно для возрастных групп или года обучения в школе), но и по его специфическим подгруппам, национальным меньшинствам, мальчикам и девочкам и т.д. Это считается показателем добросовестности составителей теста, их внимания к пользователю.

Большинство широко применяемых стандартизированных тестов составлялось ещё в докомпьютерную эпоху. Поэтому рядом с нормативами приводятся процедуры обработки результатов тестирования (с формами таблиц, графиков, технологией перевода «сырых» баллов в оценочные по шкалам субтестов и по тесту в целом). Хорошее руководство фиксирует внимание пользователя-диагноста на типичных противоречиях в ответах или особо симптоматичных для целей диагностики данных (наиболее информативных показателях).

Некоторые стандартизированные тесты имеют свои компьютерные версии и почти все компьютерные программы обработки. Обычно компьютерные версии тестов предлагаются пользователям за дополнительную оплату.

В США в большинстве штатов приняты законы об аттестации и лицензии специалистов, применяющих тесты. Обычно требуется степень доктора философии (в бывшем СССР — это уровень кандидата психологических наук), а также знакомство с практической деятельностью и удовлетворительная сдача квалификационного экзамена.

России также необходимы такие законы, ибо последние десятилетия наблюдается ситуация, которую так стараются избежать в цивилизованных странах. Методики применяют все, кому они стали доступны, а доступны они реально всем желающим и готовым заплатить за их приобретение.

Мы полностью согласны с позицией А. Анастази: «Требование, чтобы тесты использовались только достаточно квалифицированными, является первым шагом в защите индивида от неправильного использования тестов» и с тем, что для тестирования учебных достижений или профессиональной умелости нужна минимальная специальная психологическая подготовка*.

---

* Анастази А. Психологическое тестирование. С. 51.

Особо актуальным этико-методическим вопросом, который **обязательно** должен быть отражён в «Руководстве к тесту», является вопрос о степени секретности **тайны результатов** и **диагноза**. За ним стоит один из вечных принципов педагогической деятельности — «не навреди!», необходимо уважение к личности ребёнка, его гражданским правам. Об этом у нас пока только начинают говорить. Нередка ситуация, когда сугубо интимные ответы учащихся обсуждаются на педсоветах, да ещё и с вызовом родителей, которые потом «принимают свои меры».

Тест учебных достижений предусматривает, что из результатов профессионального теста может и должно стать известным педагогу-предметнику, классному руководителю, директору, родителям и, наконец, самому учащемуся. В принципе тот же вопрос относится и к неформальным тестам текущего и итогового контроля, но они на то и неформальные, что их выводы имеют очень относительную диагностическую ценность, которую легче оспаривать.

Практическое решение этого вопроса известно уже четверть века. В 1970 году в США

были изданы «Основные правила по сбору, хранению и распространению данных об учениках». В соответствии с ними при оценке пригодности достаточно согласия родителей, опекунов или других официальных представителей интересов учащихся, а при оценке личности требуется индивидуальное согласие самих учеников. Специалисты справедливо считают, что профессиональная ответственность связана не только с применением тестов и сохранением тайны ответов, но и с распространением самих тестов, с их публикацией и любыми способами тиражирования. Это полностью относится к рекламным проспектам тестов и объявлениям в педагогической прессе.

Обязательно должны быть чётко и однозначно определены в руководстве к тесту **средства профилактики негативных ситуаций тестирования**. В процессе тестирования далеко не все обследуемые горят желанием добросовестно отвечать на предлагаемые задания теста. Отказ отвечать на тест значительно искажает результаты, особенно если число отказавшихся превышает 5%. Декларированные отказы (независимо от мотивов) при массовом тестировании снижают у остальных обследуемых мотивацию работы с тестом. Особенно негативно публичные отказы сказываются на **репрезентативности** выборки теста, если инициаторами отказов выступают лидеры групп.

Попытки убеждать и уговаривать, как утверждают опытные исследователи, — ошибочны. Запуганная группа при дидактическом тестировании даёт заведомо заниженные результаты.

Способами профилактики ситуаций отказа являются, как показывает практика:

а) продуманный выбор времени тестирования;

б) определение его разумной продолжительности (с учётом того, что после 30 — 35 минут работы с тестом продуктивность решений снижается);

в) комфортность помещений, температура, отсутствие отвлекающих факторов (шум в соседней аудитории, постоянно хлопающие двери);

г) чёткое и понятное предварительное объяснение цели и задач тестирования, значимости его результатов для объективного оценивания знаний и умений учащихся;

д) внешний вид и стиль общения лица, проводящего тестирование, должны не отвлекать обследуемых, а стимулировать добросовестную работу, создавая доброжелательную атмосферу.

А. Анастази ссылается на имеющиеся данные о том, что даже включение в тест (в инструкцию для обследуемых) описания приёмов отбора заданий весьма уменьшает количество отказов из-за опасений посягательств на тайны личности*.

---

* Анастази А. Психологическое тестирование. С. 57.

Другим видом негативных ситуаций при тестировании являются **преднамеренные искажения ответов** со стороны тех, кто знает правильный ответ, но решает пошутить, позабавиться или поскорее избавиться от процедуры тестирования. Среди них встречаются любители «системного подхода». Они отмечают каждый первый (третий и т.д.) вариант ответа либо на все чётные (нечётные) вопросы дают ответ «верно», «неверно». Или на первую половину теста отвечают утвердительно, а на вторую — отрицательно.

Если при проведении тестирования педагогом-диагностом были замечены какие-либо отвлекающие «шумовые» факторы, он обязан их зафиксировать в отчёте и попытаться дать прогноз возможного влияния этих факторов на результаты тестирования. К таким факторам относятся неожиданное вторжение в аудиторию постороннего или представителя администрации, шум около аудитории, обморок или какие-либо яркие проявления болезни кого-либо из опрашиваемых (тошнота, громкий продолжительный кашель).

Особо следует отметить попытки самих тестируемых найти ответ с помощью подсказки или консультаций с другим учащимся. В этом случае необходимо **заранее в инстуциях** (для обследуемых и для диагноста) недвусмысленно оговорить порядок действий педагога-диагноста. Это может быть замечание с фиксацией «нарушителей спокойствия» и номера задания, по которому запрашивалась помощь, исключение для нарушителей этого задания из

числа оцениваемых, прекращение тестирования «нарушителей» (с удалением из аудитории и повторным тестированием в индивидуальном порядке, прекращение тестирования всей группы (если такой обмен подсказками применяет большинство).

Наряду с письменной инструкцией (или предъявляемой на экране монитора в компьютерном тесте) необходим дополнительный устный инструктаж перед сеансом тестирования, где на наиболее важные процедурные моменты должно быть обращено внимание обследуемых. Текст инструктажа, который может по ряду позиций отличаться от письменной инструкции, в целях унификации процедуры следует приводить полностью (включая рекомендации тестолога самим обследуемым с советами, на чём при работе с тем или иным субтестом стоит акцентировать внимание).

Особая группа конфликтных проблем информационно-методического обеспечения тестирования связана с допустимостью помощи опрашиваемым со стороны тех официальных лиц, которые непосредственно проводят сеанс тестирования. Эта помощь должна быть во всех типичных случаях однозначно определена в руководстве к тесту, чтобы была очевидна грань между оказываемой им помощью и санкциями в случае негативного поведения испытуемого (подсказки, шпаргалки). Надо чётко определить роль и степень полномочий самого тестолога, проводящего сеанс (американцы называют его **супервизором**). В зависимости от типа теста и целей диагностики здесь возможны вариации, но все они должны быть мотивированы психолого-педагогическими аргументами.

В методических материалах, сопровождающих известные на Западе дидактические тесты, нередко есть перечень типичных ошибок, допущенных учащимися при выполнении заданий. Особо это важно для **критериальных тестов** и диагностико-коррекционных программ. Особенно важно — для их компьютерных версий, где возможна система оперативного возврата испытуемого к базовой информации (для решения тестовой задачи) для доучивания и последующего повторного тестирования, но уже по другому варианту задания. Иногда — более сложному. Также перечни ошибок чрезвычайно важны для **педагогической валидизации** теста, выработки эффективных оперативных коррекционных мер.

Чем полнее и разностороннее в дидактическом тесте его информационно-методическое обеспечение, чем продуманнее и гуманнее составлено руководство к тесту, тем больше вероятность того, что он превратится в уважаемого и авторитетного долгожителя типа теста Бине — Симона и его не постигнет судьба бабочки-однодневки.

Оценка репрезентативности выборки апробации теста

Дидактическое тестирование, особенно на стадии разработки и доводки теста, должно подчиняться научно-методическим правилам организации и проведения эксперимента. Это относится как к критериальным, так и к нормативным дидактическим тестам.

Нормативные тесты при подготовке к массовому внедрению в учебный процесс и к публикации требуют предварительной **выверки**, необходимой для получения соотносительных норм. При этом надо стремиться к максимально репрезентативной выборке, так как может оказаться, что тест, хорошо решаемый девочками городских школ из семей служащих может плохо решаться мальчиками из школ отдалённой сельской местности.

Репрезентативностью

в социальных науках называется свойство выборочной совокупности (контингента, на котором проверялся тест) воспроизводить характеристики генеральной совокупности (контингента, для которого предназначена методика, со всеми его существенными демографическими особенностями). Достижение репрезентативности требует знания основных характеристик генеральной совокупности, чёткого описания целей обследования и того, насколько эти характеристики являются значимыми с точки зрения диагноста. К. Ингенкамп отмечает, что «информативность тестовых норм в значительной степени зависит от того, насколько велика была выборка, была ли она стандартизована или нормирована и прежде всего каким образом

она осуществлялась»*.

* Ингенкамп К. Педагогическая диагностика. М.: Педагогика, 1991. С. 51.

Стандартизированный тест для проверки знаний учащихся по какому-либо предмету должен выверяться не на особо сильных или особо слабых, а на том контингенте, который отражает статистика текущей успеваемости. В любом случае при стандартизации теста репрезентативность выборки будет определять возможную сферу его применения. Так, тест для проверки конкретного раздела знаний по высшей математике или сопромату, прошедший стандартизацию на контингенте студентов одного факультета технического вуза, может оказаться непригодным на физмате университета.

Оценка результатов нормативного теста всегда опирается на определённую **шкалу норм**. Применительно к школьным тестам шкала норм может быть признана достаточно информативной лишь в том случае, если состав выборки по городам, сельской местности, субъектам федерации, полу учащихся и т.п. пропорционален соответствующей генеральной совокупности (выборки) и отбор внутри перечисленных групп был случайным, непреднамеренным. В социологии и психологии считается, что минимально репрезентативная выборка обследуемых для проверки тестов должна включать 300–400 человек и быть максимально приближенной по демографо-биографическим характеристикам к контингенту, на который рассчитан тест (учащиеся городских или сельских школ, соотношение их по половому признаку, по классу или годам изучения предмета и т.д.). Встречались очень большие выборки, когда стандартизировались национальные тесты США. Так, рассчитанная на два дня батарея тестов способностей и достижений, интересов и темперамента проверена на выборке около 400 000 учеников IX–XII классов.

Для успешного расчёта необходимой выборки стандартизации дидактического теста (а в социологии и экспериментальной психологии — любой методики) используется ещё ряд понятий, достаточно тесно взаимосвязанных между собой.

Всякое отклонение выборки от генеральной совокупности принято называть *смещением выборки* (причины которого могут быть весьма различны, как мы видели выше). Если смещение выборки оказывается существенным, то надо будет делать дополнительные расчёты и добирать необходимый для апробации эмпирический материал либо повторно проводить обследование, либо довольствоваться тем, что сфера действия обоснованных нормативов теста окажется намного более узкой, чем первоначально планировалось. Важна также и **полнота выборки** — представленность в ней всех элементов генеральной совокупности в том структурном соотношении, которое имеется в самой генеральной совокупности (например, юношей и девушек при тестировании в школе или колледже, отличников учёбы и слабоуспевающих учащихся и т.д.).

Полнота выборки реально связана со знанием и учётом этнокультурной карты региона и соотношения половозрастных и профессиональных групп. Специфические экономико-географические условия южных регионов России, особенно Ростовской и Волгоградской областей, Краснодарского края и Ставрополья в последние годы связаны с тем, что довольно резко изменяется их этнокультурная карта.

Это, несомненно, повлияет на нормативы стандартизируемых дидактических тестов, так как дети из многих семей мигрантов в местах своего прежнего проживания получали образование нередко более низкого качества, чем то, которое они могут получить в считающихся (вполне справедливо) сильными в образовательном плане южно-российских регионах.

Отсутствие дублирования выборки — важный момент в определении её качества. При дидактическом тестировании это ситуация специфичная. Вряд ли пройдет незамеченным от экспериментаторов, если одна учебная группа по одному варианту теста будет обследована дважды, а другая— ни разу. Но подмена одного студента другим вполне возможна из-за личных опасений самих студентов или из-за нежелания педагога «подставлять» слабых учащихся, если сторонний диагност лично не

знаком с составом группы. Такое встречается даже на строго контролируемых вступительных экзаменах в вуз (в инструкциях к американским тестам оговорены различные меры предупреждения подобных ситуаций).

Точность информации по каждой единице

определяет, в конечном счёте, качество получаемых результатов. Она может быть не соблюдена: а) при невнимательной обработке (ручной или на микрокалькуляторе, компьютере) бланкового теста; б) из-за сбоев машины при компьютерном тестировании; в) из-за низкого качества бланков при традиционном способе обследования.

Адекватность выборки

для решения конкретных задач— весьма деликатный вопрос. Даже при смещениях выборки апробации (из-за слишком большой доли «сильных» учащихся) многие показатели теста — его содержательная и психологическая валидность, надёжность, могут сохранять своё значение. Диагностам, обнаружившим такое смещение, стоит задуматься: переносить процедуры тестирования на другой срок или воспользоваться тем, что есть, а потом добрать материал в соответствии с планируемой выборкой. Если процесс апробации теста затягивается на годы, то адекватность выборки может изменяться независимо от качества работы, проведённой создателем теста, если резко меняется социодемографическая ситуация. Так, в школах ряда южных городов и районов мигранты составляют до 1/4 контингента (представляя при этом широкую географию конфликтных проблем России и СНГ — от детей военных Западной группы войск, кочевавших по 4–5 школам, до темпераментных горцев, порывающихся ходить по школе с кинжалами).

Удобство работы с выборкой —

существенный организационно-методический фактор (нередко диагност идёт на изменение первоначального варианта рассчитанной выборки, что приводит к систематическим ошибкам выборки). **Объём выборки** зависит от числа признаков, относительно которых она производится и должен быть велик настолько, чтобы в каждую выделенную группировку попало достаточное количество элементов*.

---

* Гласс Дж., Стенли Дж. Статистические методы в педагогике и психологии. М.: Прогресс, 1976. С. 266.

Для лучшей ориентации в вопросах репрезентативности выборок апробации тестов педагогу-диагносту необходимо представлять, каковы бывают **основные виды выборок и методики их расчёта**. Существует несколько подходов к построению выборки — простая (случайная), квотная, систематическая вероятностная, серийная («гнездовая»).

При небольших генеральных совокупностях могут быть:

а) **простая повторная** выборка, при которой на карточки наносятся номера респондентов; карточки перемешиваются; вслепую вынимается карточка, записывается её номер; затем карточка возвращается в колоду и снова все карточки перемешиваются. При выборке 5единиц из 50 операция повторяется 5 раз;

б) **простая бесповторная** выборка, при которой отобранные карточки откладываются.

Независимо от вида выборки возможны и естественны **ошибки репрезентативности**.

Для простой вероятностной выборки требуется перечень всех единиц генеральной совокупности, что пригодно для небольших исследований в рамках школы, малого города или небольшого сельского райотдела образования.

Квотная выборка

представляется как модель генеральной совокупности в виде квот (пропорций) распределения изучаемых признаков (сколько лиц с какими характеристиками надо опросить). **Систематическая вероятностная выборка** — это упрощённый вариант вероятностного отбора. В основу выборки, как правило, кладут различные алфавитные списки, картотеки и т.п. Отбор

единиц осуществляется через один и тот же интервал из исходного алфавитного или пронумерованного списка возможных кандидатов в обследуемые.

Систематическая выборка является экономным и удобным способом формирования выборочной совокупности. Следует учитывать возможность систематического распределения в списках единиц различного типа, повторяемости в их распределении, которая может совпадать с величиной **интервала отбора**, равной К (пример смещения — ведомость на зарплату, составленная в порядке её уменьшения или список учащихся, составленный по их учебному рейтингу, — как раньше было в царской России и теперь восстанавливается в ряде учебных заведений, перешедших на тестовую систему контроля знаний). Эту выборку лучше использовать при однородной генеральной совокупности.

Серийная («гнездовая») выборка

более сложна, но обычно — и более точна. Единицы отбора — это статистические серии, т.е. совокупности статистически различимых единиц (семья, учебная группа, школа определённого типа). Отобранные в выборку серии подвергаются сплошному обследованию. Серийная выборка может организовываться по схеме простой, случайной и систематической выборки или формироваться после предварительного районирования генеральной совокупности. При этой выборке обычно имеет место занижение дисперсии изучаемого признака из-за определённого сходства единиц в сериях. Она применима при обследовании целых школ в районе, классов — в школе. Нередко при этом получается избыточная информация, но это органический порок данного вида выборки.

Серийный отбор

будет тем репрезентативнее, чем меньше степень колебаний серийных средних, измеряемая величиной их дисперсии. Если тест создаётся самими учителями, то к расчёту выборки лучше привлекать опытных социологов или вузовских учёных.

Расчёты эти весьма важны для диагноста. Так, все показатели теста SAT — одного из популярнейших и долгоживущих в США, построены на среднем значении и стандартном отклонении данных почти 11 000 абитуриентов, проходивших в США этот тест ещё в 1941 году. Они составили **референтную группу** (своего рода эталон), относительно которой проводилось шкалирование всех последующих форм и модификаций этого широко применяемого теста.